DELAYER PAYS PRINCIPLE
Examining Congestion Pricing with Compensation

By David Levinson and Peter Rafferty
Assistant Professor Graduate Student
(Corresponding Author)

Department of Civil Engineering
University of Minnesota
500 Pillsbury Drive SE
Minneapolis, MN 55455

levin031@umn.edu raff0040@umn.edu
Work: 612.625.6354 608.270.5384
Fax: 612.626.7750 608.274.2026

Submitted for 2003 Transportation Research Board Conference
November 7, 2002

Word Count:
5,609 words + 7 tables/figures x 250 = 7,359

ABSTRACT
Despite its virtues, congestion pricing has yet to be widely adopted. This paper explores the issues of equity and use of toll revenue and several possible alternatives. The equity and efficiency problems of conventional (uncompensated) congestion pricing are outlined. Then, several alternatives are discussed and developed. A new compensation mechanism is developed, called the "delayer pays" principle. This principle ensures that those who are undelayed but delay others pay a toll to compensate those who are delayed. We evaluate the effectiveness of this idea by simulating alternative tolling schemes and evaluating the results across several measures, including delay, social cost, consumer surplus, and equity. Different tolling schemes can satisfy widely varying policy objectives, thus this principle is applicable in diverse situations. Such a system is viable and can eliminate some common hurdles of congestion pricing – while remaining revenue neutral.

INTRODUCTION

Congestion pricing continues to face barriers to adoption, yet the basic theory is uncomplicated. Vickrey (*1*) introduced a simple bottleneck model that illustrated how pricing at a roadway bottleneck is an effective way to eliminate delay. He showed that tolls charged to drivers could spread the demand evenly through the rush period to reduce or eliminate delay, while maintaining the same throughput at the bottleneck.

Much theoretical development has occurred since Vickrey's paper. Twenty years later, Small et al. (*2*) devoted a good portion of their book *Road Work* to congestion pricing and its policy implications. They clearly affirm that congestion pricing theory is well developed and accepted. In fact, test cases and empirical evidence were already available at that time. Yet implementation remained scarce because of political hurdles. In 1989, only Singapore had any form of congestion pricing in place. Of course, the political environment there is different from the United States. In the United States, road use is largely free of tolls, though gas taxes are charged, and additional taxes – as road pricing is perceived to be – are unpopular.

While traffic is very light, little interaction occurs between vehicles and no congestion externality affects the cost of a trip. When traffic is heavier and an additional vehicle imposes added delay on other drivers, the marginal cost of a trip is greater than the average cost. Because drivers base their decision to make a trip on perceived average cost, the untolled equilibrium occurs at a greater traffic volume than if the decisions were based on the marginal cost. A goal of congestion pricing is to adjust the personal cost of a trip in congested conditions such that equilibrium occurs at the lower volume where the marginal cost curve intersects the demand curve. This increased cost internalizes the congestion externality.

Many second-best variations of this basic congestion pricing idea have been proposed – and some implemented – that hope to improve economic efficiency, given practical considerations. Braid (*3*) and Verhoef et al. (*4*) both explored situations with untolled alternative routes. Others have proposed rationing or reservation pricing as second-best solutions, and several metropolitan areas have implemented variations. Fielding and Klien (*5*) proposed High Occupancy Toll (HOT) Lanes, which now operate in several United States cities. Daganzo and Garcia (*6*) developed a scheme that combines rationing and reservation pricing for the bottleneck model. They showed that this second-best solution could reduce user cost while improving Pareto efficiency. As is common in congestion pricing writing, their paper concludes by listing several practical and technical questions that need answering.

These questions typically wonder at what to do with cheaters and uninstrumented vehicles; how will alternate untolled routes come into play; what are the equity implications – both for the rich vs. poor users, and for the areas in a metropolitan region with and without toll roads; and how should the tolling authority use the revenue?

What to do with the revenue is another critical question that needs to be answered before toll roads will become more widely adopted. One possibility is compensating the delayed. This paper investigates the issue of compensation and several possible alternatives. First, the equity and efficiency problem of conventional (uncompensated) congestion pricing is outlined. Next, a new compensation mechanism is suggested, called the "delayer pays" principle. These alternatives are in contrast with the efficiency arguments put forward about marginal cost pricing presented in most research on the subject.

Within the development of the delayer pays principle, we explicitly quantify the full marginal cost. An overlooked aspect of the marginal cost in traffic congestion is that the delay externality caused to other users may persist beyond the time a given vehicle is present in a

queue. Properly pricing this is crucial for efficient and equitable tolling. In the delayer pays scheme there is a transfer of resources, not just a collection of revenue. This scheme is also different from those that offer reduced tolls, vouchers, or other compensation because the two-part dynamic toll is directly linked to the congestion externality and marginal costs. Drivers will pay the price that corresponds to the full marginal cost (delay caused to others), but they will also collect the corresponding compensation for the delay they experience, which would not exist but for the presence of others. In this scheme, it is feasible for the tolling authority to act only as a transfer agent and collect no net revenue.

Presented last is a further exploration of delayer pays pricing in a broader scope. We evaluate the effectiveness of this idea by simulating many different tolling schemes and evaluating the results across several measures. Issues of modeling, equilibrium, and policy are discussed along the way. This effort offers solutions to varied policy objectives and brings us closer to practical implementation of congestion pricing.

DELAYER PAYS PRINCIPLE

Previous marginal cost conceptions did not consider the full consequences of delay caused by a vehicle joining a queue. For example, Arnott et al. (*7*) describe the Bottleneck Model and state, "The marginal cost of a driver is independent of when she departs [from home]." However, consider a driver choosing whether to depart early and reach the bottleneck early in the queue or to depart home later and pass through the bottleneck toward the end of the queue period. The delay and the entire travel cost to the driver could be the same in either case, but the marginal cost is very different. In the first case, the driver imposes an incremental delay on every driver to arrive behind her during the queue period. In the second case, only the small number of drivers to arrive after her at the end of the queue will face the additional delay. A vehicle's presence earlier in the queue may have a much longer reverberation. The full marginal cost of a vehicle depends on how many other vehicles there are and when each vehicle arrives.

Charging the full marginal cost and paying people proportional to their delay would produce the result shown in Figure 1. This figure is a decomposed queuing diagram, and it illustrates a simple example of nine vehicles arriving in nine seconds, thus forming a queue which then dissipates. Vehicle one arrives and then departs as vehicle two arrives. Vehicle three is delayed by vehicle two and delays vehicles four and five. By the time vehicle five passes the queue, vehicle nine is already waiting, and so on. This representation also illustrates that more money is paid in than paid out. The discrepancy is because eliminating a vehicle will sharply reduce delay, but to the delayed vehicle, it matters not which vehicle ahead is eliminated, any one of them will reduce delay. With congested facilities, additional vehicles raise the average travel cost for everybody – thus the marginal cost always exceeds average cost. So using the full marginal cost accounting will generate surpluses. This can be described mathematically with the equations and description given in Table 1. The incomplete marginal cost corresponds to the queue at the time a vehicle departs the queue; the full marginal cost also accounts for the vehicles arriving in the queue after this time. The reimbursement income corresponds to the delay caused to a vehicle by others before it.

If people vary in their values of time, people with a high value of time may not be fully compensated, while those with a low value of time would get more dollars back than the value of the time they lost. This may induce more travel by clever people with low values of time trying to swindle the system. Without a base toll in place, a person arriving at the end of a queuing

period is delayed more than they delay others and could receive a net compensation. However, a nominal base toll eliminates this possibility while funding administration, operation, and maintenance costs.

Moreover, the system would send price signals back to drivers, who would then adjust their departure times in some fashion, thus smoothing out the demand. A new, less peaked, arrival pattern would result. Therefore, after equilibration between price and demand, the system would have a lower price and lower net turnover.

Strictly speaking, the correct charge is unknown until some time after the driver exits (the front) of the queue, but some approximations could be made. The charge depends not only on how many vehicles were behind the driver at the time the driver exits, but also on how many vehicles are behind those vehicles – that is on how much delay that vehicle actually caused. We can post the expected price on a variable message sign just before the bottleneck. This will not be strictly accurate, as the mainline flow may suddenly spike upward, or the off-ramp may suddenly get more traffic. Nevertheless, with experience, the forecasting system would become increasingly reliable.

The delayer pays scheme is a straight-forward solution to "what to do with congestion pricing revenue" – return it directly to those who were delayed almost instantly. The system can be perfectly revenue neutral, stay within the roadway sector, and be economically efficient. Overall, the amount of revenue collected could equal the amount distributed. However, those who delay others the most pay the most, while those who are delayed more than they imposed delay on others are compensated for their delay.

Policy Objectives and Measures of Effectiveness

Defining an objective function can be a complicated task. Because the policy objectives of congestion pricing will vary among jurisdictions, we will leave the definition open-ended. Nonetheless, we wish to explore the feasibility of this type of tolling scheme, so we have identified seven measures of effectiveness. These measures may enter the objective function in various combinations and with various weights.

1. Total Delay is the first measure. This is the total queuing time at the bottleneck for all vehicles. It is the area between the arrival and departure curves on a standard queuing diagram. This value should be no greater than the untolled condition.

2. Schedule Delay is the sum of all early and late arrival penalties arising from drivers missing their desired arrival time. A value of time is assigned to convert this measure to dollars.

3. Total Toll is the net toll from the user's perspective. If this value is zero, then the tolling authority has no net gain, but they do if this is positive. A certain criterion is that this value must be greater than zero, or the tolling authority loses money. While net payment to drivers may reduce delay, it is an unlikely policy decision. The application of a base or constant toll can ensure that no driver can profit.

4. User Cost is the dollar sum of the first three items: total delay, schedule delay, and total toll. Another objective may be to improve this value, though if it is unchanged, other measures may still improve.

5. Social Cost considers tolls – both positive and negative – as purely transfers between agents within the system. Therefore, the social cost is the sum of only the total delay and the schedule delay. This is a key value to minimize because it represents economic inefficiency arising from congestion externalities.

6. Equity among users is measured by the Gini coefficient associated with a Lorenz curve. This measure is not about the spatial equity problem in a metropolitan area when tolls are applied to an isolated road; that is a broader policy issue and is not addressed here. A Lorenz curve is developed which represents how the share of the cost is spread among the population. The Gini coefficient is a measure of the deviation from perfect equity. A coefficient equal to one is perfect inequity (one person is paying for all); a coefficient of zero is perfect equity.

7. Consumer Surplus is the seventh measure. This is estimated by evaluating the logarithm of the denominator in the choice equation (defined in the next section)

$$CS_k = \ln \sum_{j=1}^{12} \exp\left(-C_j\right)$$

where CS_k is the consumer surplus for tolling scheme k, j indicates the first through twelfth time intervals for the given day, and C is the cost of the trip for the time period. This value is also known as the log-sum (*8*).

Another consideration is how closely the tolls compare to the theoretical marginal cost pricing. How closely does the positive toll pattern matches the theoretical right triangle shape; how much is each group paying for the delay caused; and how much are they reimbursed for delay suffered?

The possible combinations of these measures into an objective function are limitless. Only a few are discussed here, but other objectives are equally suitable

DELAYER PAYS MODEL

The delayer pays model is based on the bottleneck framework – and its extensions – of many earlier efforts (e.g., *1, 6, 9*). A number of motorists desire to pass a bottleneck at a certain time during the morning commute. Departure time decisions are modeled with a multinomial logit model with a random utility component. Values of time are assumed the same for all travelers.

Figure 2 illustrates a characteristic queuing diagram for a large number of vehicles. Time is measured on the horizontal axis, number of vehicles on the vertical axis. The slope of the arrival curve (the upper curve) represents the rate at which vehicles are arriving at a bottleneck; two different arrival rates are shown in this figure. The slope of the departure curve (the lower curve) is the rate at which the facility serves traffic. The departure curve is the same as the arrival curve unless arrivals exceed capacity or a queue is already present. The space between the two lines represents the delay in a queue. The vertical distance between the arrival and departure curves at any time is the number of vehicles in the queue. The horizontal distance is the time spent in the queue by any vehicle. This is a standard representation of traffic flow at a bottleneck.

In the absence of vehicle n, every vehicle arriving after it saves the time that vehicle n took to pass the bottleneck. Assuming a single-lane bottleneck with a capacity of 1800 vehicles per hour, the service time per vehicle is two seconds. The heavy line along the arrival curve in

the figure represents the delay externality caused by vehicle n. The height of this line – the number of vehicles from n to the last queued vehicle – multiplied by two seconds is the toll that ought to be charged. The determination of this value extends in time beyond t_d, the time that vehicle n departs the queue. The heavy horizontal line represents the delay experienced by vehicle n, caused in part by each vehicle arriving before vehicle n. Therefore, the heavy horizontal line represents the reimbursement to vehicle n.

This scheme raises important questions. The shape of the full marginal cost toll in time is of chief concern in practice. As in Figure 1, this toll jumps from zero to its maximum value for the first vehicle in the queue, and then reduces again to zero over the duration of the queue. The implication, of course, is that for a very large facility the first driver in the queue would be charged a lot of money to pay for the holdup caused to the possibly thousands of vehicles to come behind. Can (or should) this be rectified to enable implementation? Another question is whether welfare gain can be realized with zero net revenue for the tolling agency. Tolls are collected for the delay caused, but the money is allowed to be returned in part, in full, or even in excess, for the delay experienced. The tolling agency may act only as a transaction manager for the delayers paying the delayed.

Methodology

This investigation uses a hypothetical bottleneck section to represent a capacity constraint. The number of lanes approaching the bottleneck is two or more lanes, but the departure from the bottleneck is just one lane. The service time assumption is two seconds per vehicle, corresponding to a typical maximum throughput of 1800 vehicles per hour per lane.

During a morning commute, 1200 vehicles ideally wish to pass this bottleneck at 8:00 AM, and they wish to do so with minimal delay. A driver passing earlier than this would arrive at work earlier than necessary and would be foregoing time that could have been spent at home or doing something they feel is a better use of their time. A driver passing the bottleneck later than 8:00 AM will arrive at work later than desired and must deal with the associated penalties. Not only are they late for work and have lost that time, but also they may have to make up that time later. These early and late penalties are also referred to as schedule delay.

The entire cost, or disutility, of a trip for each user comprises six components: (1) early arrival penalty, (2) late arrival penalty, (3) delay penalty, (4) positive toll, (5) negative toll, and (6) base toll. The values of time among all drivers are the same. A typical value of time for motorists, commonly used in benefit-cost accounting for example, is 9 dollars per hour, corresponding to the 15 cents per minute used in this model.

The early arrival penalty decreases linearly with the time the vehicle passes the bottleneck as time approaches 8:00 AM; it is zero otherwise. The tradeoff assumed for this time is 10 cents per minute, which reflects the different values of time at home and time spent at the office early. The late arrival penalty increases linearly with the time the vehicle passes the bottleneck after 8:00 AM, and is zero if they pass before then. Arriving late has a cost of 20 cents per minute.

The extra time the trip takes due to congestion is the delay component. In this model, the delay occurs at the bottleneck – in real networks the principle in unchanged, but the situation is more complicated. The value of extra in-vehicle time is the same as the value of time, 15 cents per minute. The travel time and operating costs for the entire trip in the absence of congestion

are not included in this model because they are an underlying fixed cost the user has chosen on a long-term basis by the location of their residence and employment.

The positive toll is the full marginal cost of congestion caused to others. This is calculated as the total delay that would save in the absence of the vehicle. As before, the cost of delay is 15 cents per minute.

The negative toll reimburses drivers for the delay they experience. The objective of this tolling scheme is not to generate revenue for the tolling agency, but to internalize the external costs of congestion. Our hope for this compensation aspect is that it has public and political appeal. One pays for the congestion caused to others, and one is paid for the congestion suffered from others. Drivers will not be paid so much as to attract profiteers, for the cost of time and operating a vehicle outweighs the negative toll remuneration. A base toll can also offset this potentiality. In this model, the negative toll is set at 15 cents per minute.

The implications of the positive and negative tolls are best illustrated by considering the first and last vehicles arriving in a queue. The first vehicle experiences no delay but causes some small delay to every vehicle queuing from that point until the dissipation of the queue, which again obviously occurs after the given vehicle has departed from the bottleneck. If a separate toll were applied to every vehicle, then this vehicle would be paying the maximum toll.

Vehicles arriving near the time of queue dissipation will not cause much delay, and therefore have a very small positive toll, but still collect the negative tolls for the time they are delayed. This results in a net money flow to this vehicle. As mentioned before, tangible income is unlikely. A base toll should only be set high enough to ensure that net tolling is greater than zero. Regardless of its value, it represents another fixed cost so does not affect time interval choice and demand patterns.

Driver Time Interval Choice

Drivers choose the time they will arrive at the bottleneck according to their perceived disutility for traveling at that time. Rather than determine the utility for each of the 1200 vehicles, a utility is calculated and averaged for each five-minute period surrounding the ideal passage time of 8:00 AM. For this investigation, 12 possible 5-minute time slots are presented to the drivers, from 7:20 AM to 8:20 AM. The 12 periods provide a wide range of utilities, and the five-minute increments are small enough to provide an approximation of a continuum, but large enough to encompass many vehicles and ease computation. In reality, one cannot expect drivers to gauge their arrival time much more precisely than a five-minute window. In addition, the five-minute period matches a typical data collection time increment on freeways and could be the time increment used in practical congestion pricing applications. The beginning and ending times were established such that the utilities associated with travel at those times are approximately equal. At the beginning and end of the peak hour only the early and late arrival penalties are usually in effect.

How many drivers choose each of the 5-minute periods is determined throug a random utility multinomial logit choice model. The underlying cost equation is:

$$C_i = E*t_{ei} + L*t_{li} + D*t_{di} + P_i - N_i \qquad i = 1,\ldots,12$$

where,

C_i is the cost for time interval i
E is the cost/minute of passing the bottleneck early
t_{ei} is the average time before 8:00 AM that group i passes the bottleneck (minutes)
L is the cost per minute of passing the bottleneck late
t_{li} is the average time after 8:00 AM that group i passes the bottleneck (minutes)
D is the cost per minute of delay
t_{di} is the average time spent in the queue for group i (minutes)
P_i is the positive toll for group i
N_i is the negative toll (reimbursement) for group i

A term for a fixed base toll is not shown here. The number of vehicles choosing time period i is therefore:

$$V_i = 1200\left(\frac{\exp(-C_i)}{\sum_{j=1}^{12}\exp(-C_j)}\right)$$

where V is the number of vehicles, i represents the time interval (one through twelve) in question, j indicates the first through twelfth time intervals for the given day, and C is the cost of the trip for the time period.

The solutions to this arise through an iterative process. Drivers will make their decisions based on the 12 choices presented to them. The utility of these 12 choices in turn depends on the decisions of the drivers and the resulting delay. Therefore, the drivers choose their arrival time based on "yesterday's" results. Each component of the cost is averaged over all vehicles in the 5-minute interval, and the sum is the information presented to the decision makers on subsequent days.

The headways within each 5-minute interval are assumed constant. If the average headway is less than two seconds, a queue forms. This occurs if more than 150 of the 1200 vehicles choose any interval.

Positive Tolling Schemes

Many positive tolling schemes were tested and evaluated, and the theoretical triangle shape shown in Figure 1 remains a subset of these alternatives. Figure 3 illustrates the possibilities. The simplicity of the rectilinear shape allows for definition by just four parameters. Certainly, an optimization routine could sort through independent tolls among all intervals, but such a resulting scheme would be difficult to implement in the field and confusing for drivers. The only constraint on the shape is $a < x < b$. We allow the position of a to vary among the first eleven time intervals. In the figure, a is in position 3, and x is in position 6, and b is in position 11. Lastly, the peak toll is controlled by the parameter y. For this exercise, this parameter varies from \$0.00 (a no-toll condition) to \$2.40 in eight \$0.30 steps. In the figure, y is in position 5, or \$1.50. The tolls in the intervals before and after this peak toll are linear interpolations based on the four parameters. In Figure 3, {a,b,x,y} = {3,11,6,5}. There are 1,848 possible tolling

schemes to evaluate using a variant of a grid search optimization. The toll free condition remains the baseline for comparison.

EVALUATION OF TOLLING SCHEMES

With no tolling, the vehicles assume an expected arrival pattern. A queue is present from 7:45 to about 8:10, and the total delay is 52.0 hours. The cost of the trips in the early and very late periods is controlled solely by the schedule delay.

It is because of the conflicting goals that we developed 1,848 trials of all triangular shapes and sizes of positive tolling schemes to compare to the untolled condition. However, it is cumbersome to visualize the relationship between the seven measures of effectiveness at the same time. It is much easier to plot the results for each of the 21 possible pairs of measures to see where the untolled condition lies. As much as each relationship warrants at least a paragraph of discussion, just two are reproduced here.

In Figures 4 and 5, the unique mark represents the untolled condition. The "tails" are those alternatives with a very small positive toll, but with the negative toll in full effect. Travelers in these cases pay little but are reimbursed for their time in the queue. In each figure, there are many points showing that the two axis measures can be improved over the untolled condition simultaneously. In Figure 5, any tolling scheme represented by a point below and left of the mark improves both total user cost and total delay.

Are there any scenarios such that *all seven* measures are improved? The answer is yes, but only two. Figure 6 illustrates the positive toll patterns for these two scenarios. Note that they are both of the form of the theoretical marginal cost triangle. In addition to improving the seven measures discussed, each of these scenarios improves the utility for all but about 8% of drivers. These solutions are therefore nearly Pareto improving strategies; and those 8% that are worse off face only a five-cent increase in their total trip cost.

Alternatively, an agency can easily select a tolling scheme that best suits their policy objectives.

There exists a tolling scheme that effectively eliminates delay. While reducing delay is a good thing, this scenario sacrifices user cost, schedule delay, and consumer surplus. Maximizing equity, or minimizing the Gini coefficient, may be another policy objective. However, the scheme that maximizes equity also has a very high user cost. A third objective may be to reduce social cost. Unfortunately, the scheme that does that best has a very high user cost and the queue is split into two – one before the peak toll and one after. Regarding consumer surplus, the best way to maximize this is to not charge a positive toll while continuing to reimburse motorists for their delay – the "tails" in Figures 4 and 5 – very expensive for an agency.

These single measure objectives are unlikely policies, but they do illustrate the tradeoffs among the conflicting objectives. As shown earlier, two scenarios do improve all seven measures, but there remain infinite middle-ground scenarios. Figure 6 also shows two more possibilities.

The first alternative satisfies the objective of maximizing social welfare – or minimizing social cost – while ensuring that user cost does not worsen and that the total toll is positive. If a base toll is not included, 15 scenarios satisfy these criteria. With a $0.25 base toll, that number rises to 30. This improves social welfare by about 12% and reduces delay by about 40%. The other measures for this solution are all within 5% of the untolled condition.

The last scheme minimizes delay and maximizes social welfare, with similar constraints. This solution yields a 53% delay reduction, an 11% social cost reduction, a 19% Gini coefficient improvement, and a 16% increase in consumer surplus. Schedule delay and user cost are nearly unchanged (within 1%) from the untolled condition, and the tolling authority collects a small net revenue. This second-best solution is very promising because it shows that a simple tolling scheme – with compensation – can improve both efficiency and equity.

CONCLUDING REMARKS

Equity and efficiency form the two pillars on which transportation decisions should be made. However, determining what is efficient, much less what is equitable, is far from simple. Who owns the right to travel on the roadway? Currently the system is first-come first-serve. Unfortunately the conventional marginal cost pricing approach often ignores traffic dynamics and tends to treat time in discrete blocks rather than continuously. How significant a problem this is depends on the conditions of the case. The delayer pays scheme outlined in this paper implies everyone has a right to free-flow, and the individuals who deny that right to others are the ones who should pay. So is delayer pays a good idea? This depends on answers to two questions:

- Empirical question - What will be the magnitude of cheating/gaming the system?
- Technical question - What is the cost of the added data collection and toll redistribution?

Traffic manifests high transaction costs, no property rights, and little bargaining, perhaps explaining the lack of efficient outcomes. Electronic tolling obviates transaction costs, and we can consider at least two extreme alternatives regarding the initial distribution of rights:

- Everyone has the right to free (unpriced) travel.
- Everyone has the right to freeflow (undelayed) travel.

If everyone has the right to free (no monetary cost) travel, then the mechanism for more efficient travel requires the delayed to pay the delayers not to delay (a congestion prevention mechanism), or the delayed will continue to suffer congestion. Alternatively, if everyone has the right to freeflow (undelayed) travel, then the burden is on the delayers to compensate the delayed (a congestion damages mechanism). We have demonstrated that the delayer pays principle – with compensation – can lead to efficient outcomes.

There are also several key philosophical questions that need to be addressed. These very much parallel the fundamental question of whether people should be guaranteed equality of opportunity or equality of outcome. Congestion externalities require two actors: the delayer and the delayed. If both parties have equal opportunity to arrive, then one should not compensate the other. However, if we want to guarantee an equal outcome in terms of a combination of time and money, those who save time should pay more money and those who spend more time should be paid by those causing their delay.

Congestion pricing generates revenue that can substitute for conventional transportation financing (such as the gas tax). Few argue against substitution, as it makes sense as a demand management measure. However, what to do with excess congestion pricing revenue has been a hurdle for its adoption. In the absence of private roads, this is a political problem. Suggestions range from the government keeping the money, to building more roads, to providing transit, to

compensating the poor (redistributing the money by income class). There is a clear alternative however that is fair, returning the excess congestion pricing revenue to those who are congested, in the form of cash or credits, with a nominal base toll in place to stave off gaming of the system.

This paper presented the results of extending the delayer pays framework to an experimental condition where 1200 drivers face a morning commute bottleneck. It is clear that the marginal congestion cost had until recently been incompletely interpreted, but now that it is fully identified, the realization of marginal cost congestion pricing can be studied further. There are substantial practical considerations that require further thought regarding the shape of the long-range marginal cost toll when extended to 1200 or several thousand vehicles. It is reassuring to see that diverse objectives can be met simultaneously. This is a key finding, with implications for further welfare and equity improvement. We have also shown that correcting the congestion externality is tenable without making other measures worse for drivers. This is another step closer to more efficient road financing.

Further and more focused research should be made around those tolling schemes that demonstrate meeting the objectives of accurately pricing marginal congestion cost, reducing delay, balancing positive and negative tolls, maintaining overall user cost, and improving social welfare, equity, and consumer surplus.

REFERENCES

1. Vickrey, W. Congestion Theory and Transport Investment. *American Economic Review*, Vol. 59, 1969, pp. 251-260.

2. Small, K., C. Winston, and C. Evans. *Road Work: A New Highway Pricing & Investment Policy*. Brookings Institution, Washington, DC, 1989.

3. Braid, R. M. Peak-Load Pricing of a Transportation Route with an Unpriced Substitute. *Journal of Urban Economics*, Vol. 40, 1996, pp. 179-197.

4. Verhoef, E., P. Nijkamp, and P. Rietveld. Second-Best Congestion Pricing: The Case of an Untolled Alternative. *Journal of Urban Economics*, Vol. 40, 1996, pp. 279-302.

5. Fielding, G. J., and D. B. Klein. How to Franchise Highways. *Journal of Transport Economics and Policy*, May 1993, pp. 113-130.

6. Daganzo, C. F. and R. Garcia. A Pareto improving strategy for the time-dependent morning commute problem. *Transportation Science*, Vol. 34, 2000, pp. 303-312.

7. Arnott, R., A. de Palma, and R. Lindsey. Recent Developments in the Bottleneck Model. In *Road Pricing, Traffic Congestion, and the Environment*. Edward Elgar Publishing, Inc., Northampon, MA, 1998.

8. Ben-Akiva, M. and S. Lerman. *Discrete Choice Analysis*. The MIT Press, Cambridge, MA, 1985.

9. Arnott, R., A. de Palma, and R. Lindsey. A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. *American Economic Review*, Vol. 83, No. 1, 1993, pp. 161-179.

LIST OF TABLES AND FIGURES

TABLE 1 Mathematical Model of Delayer Pays Compensation Schemes

Cost and Income Variables	Expression
S_v = Own cost	$S_v = A_v - D_v$
$T_{[\ ]}$ = Total cost [for arrival pattern containing vehicles in bracket]	$T_{[\ ]} = \sum_{[\ ]} S_v$
J_v = Incomplete marginal cost	$J_v = Q(D_v) - 1$
M_v = Full marginal cost	$M_v = T_{[1\text{--}V]} - T_{[1\ \text{--}\ v\text{-}1, v+1\ \text{--}\ V]} - S_v$
R_v = Reimbursement income	$R_v = S_v / \mu$
N_v = Net income	Incomplete marginal cost $N_v = J_v - R_v$ Full marginal cost $N_v = M_v - R_v$

Notes: Subscript $_v$ denotes vehicle v. A_v = Arrival time (at back of queue). D_v = Departure time (from front of queue). $Q(t)$ = Number of vehicles in queue at time 't'. μ = Service time (headway between vehicles departing queue).

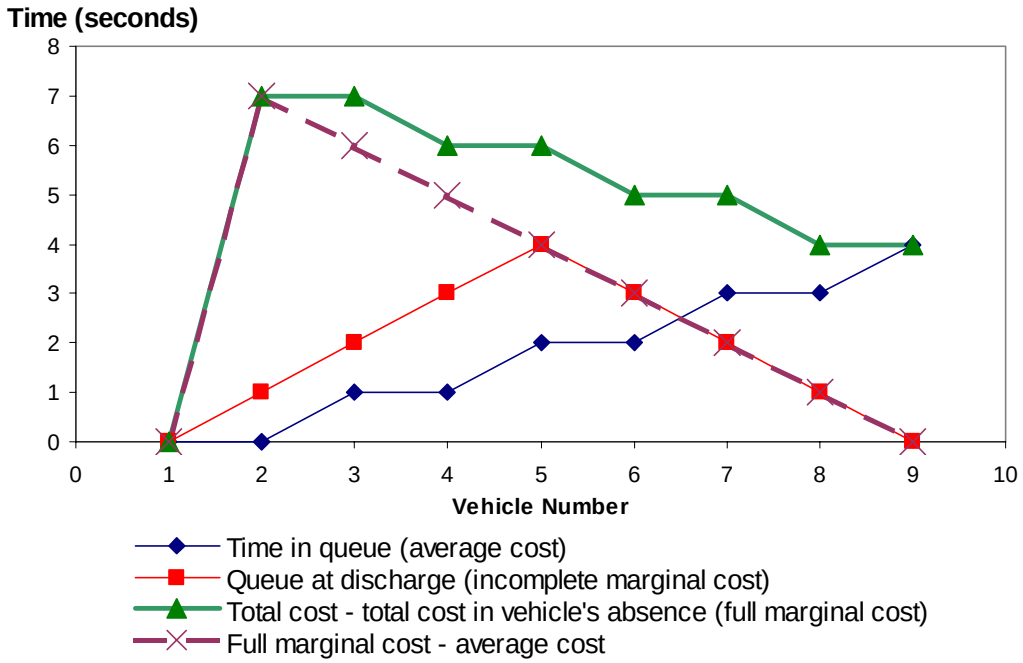

FIGURE 1 Average and Marginal Effects of Delayer Pays Principle (for nine vehicles)

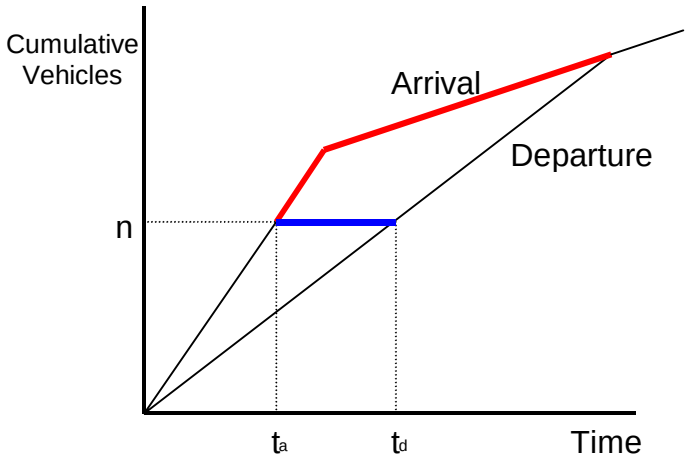

FIGURE 2 General Queuing Diagram

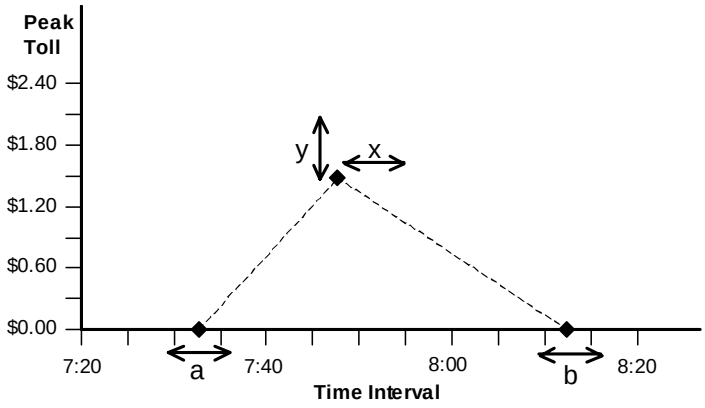

FIGURE 3 Positive Toll Alternatives

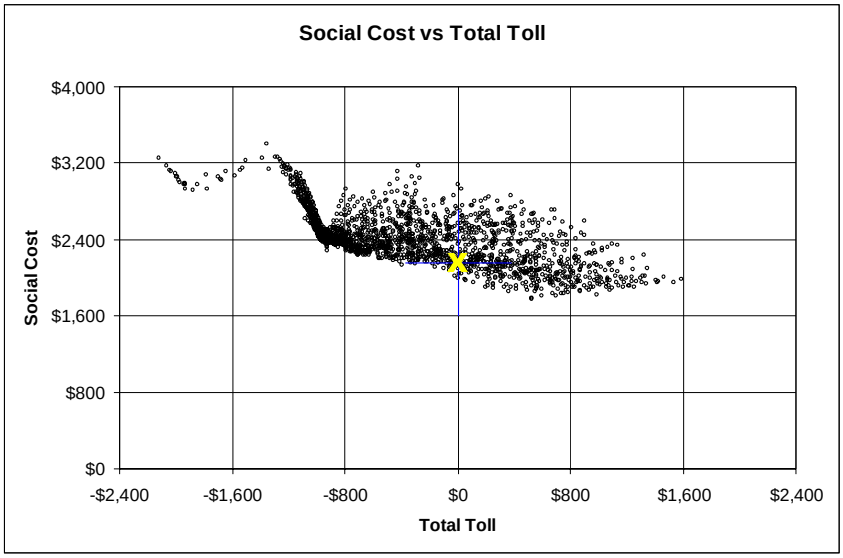

FIGURE 4 Social Cost vs. Total Toll

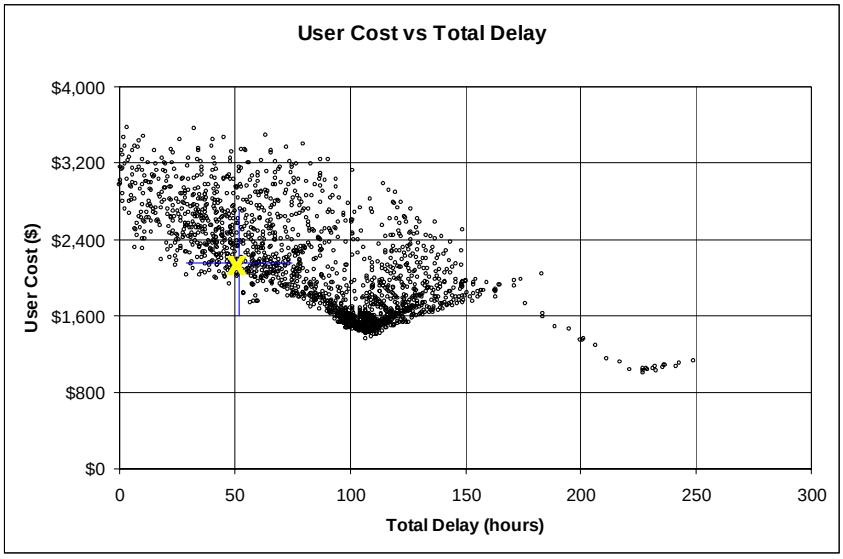

FIGURE 5 User Cost vs. Delay

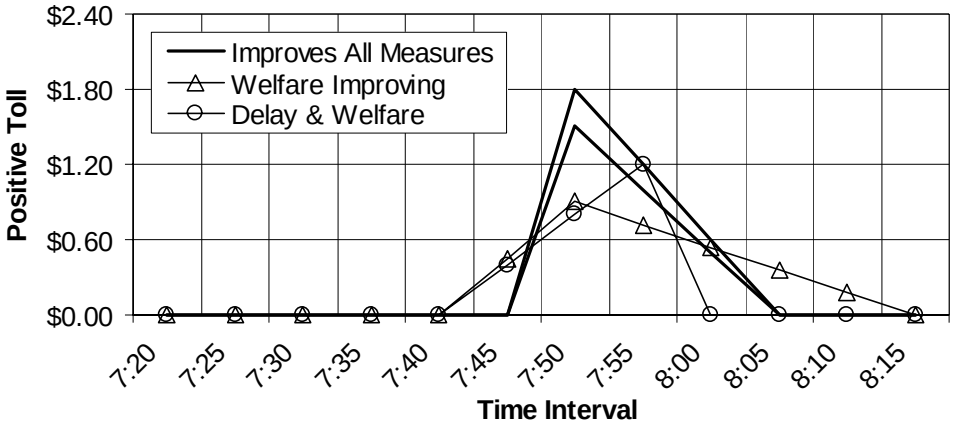

FIGURE 6 Alternative Tolling Schemes